I/O Hardware

Basic I/O Hardware

"Serial" Systems

- **Cray C90**

Parallel Systems

- **Intel Paragon**
- **IBM SP2**
- **Meiko CS-2**
- **Cray T3D**

Parallel I/O Hardware

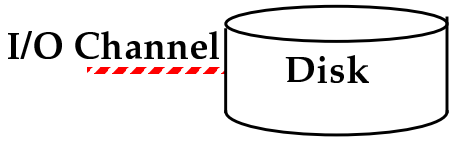
Disks

Attaching disks to compute nodes

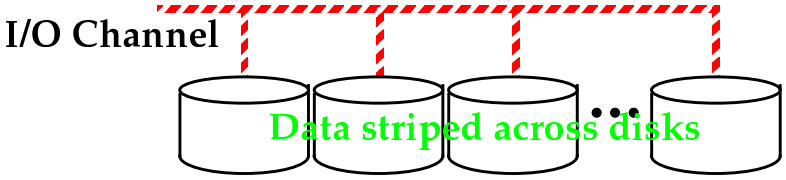
- **Directly attached disks**
- **Software arrays/RAIDs**
- **Network attached disks**
- **I/O node(s)**

Disks

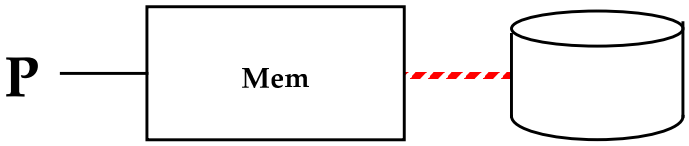
Disks



Hardware disk array/RAID



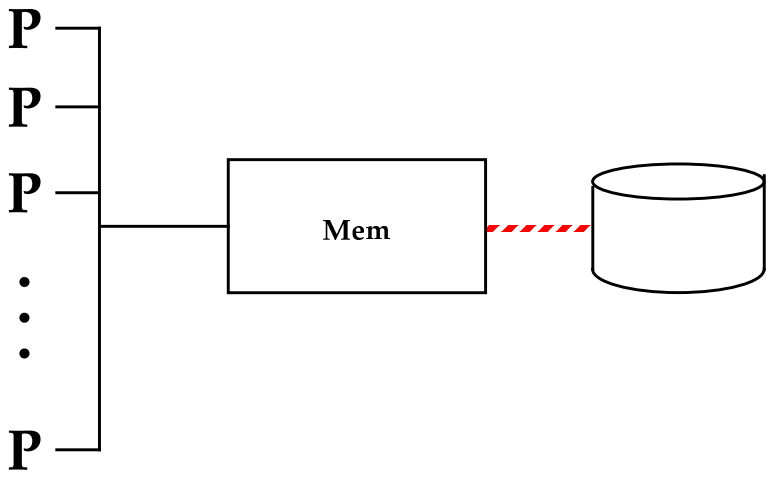
Directly Attached Disks (Workstations)



Disks attached directly to nodes

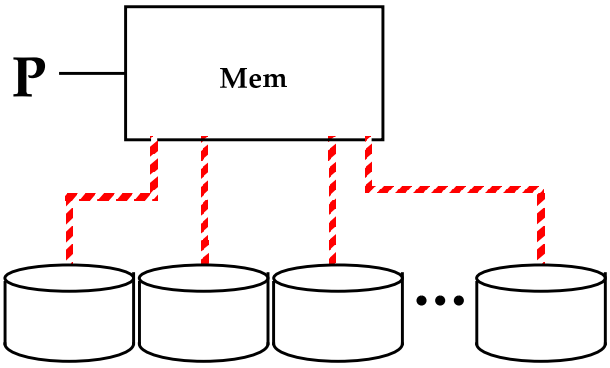
- Only single processor can access disk
- Disk attached with SCSI or similar interface
- Other processors must request data on disk from attached processor
- Remember, the disk could be a hardware array

Directly Attached Disks (SMPs)



Like directly attached disks, but all processors in the "box" can directly access the disk

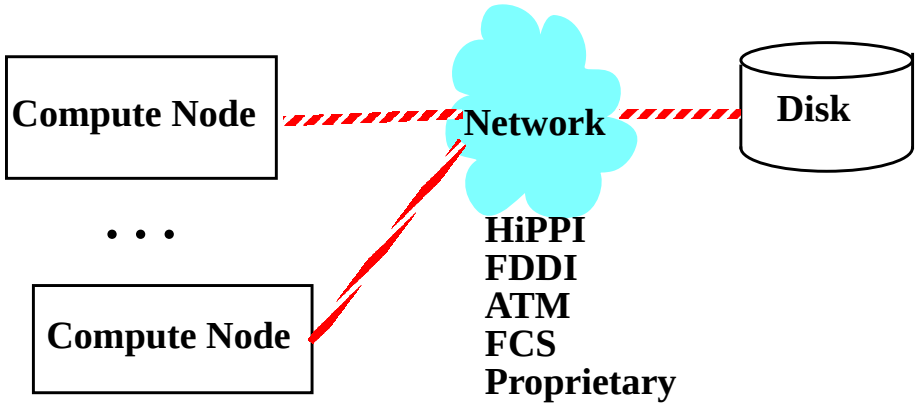
Software Arrays/RAIDS



Multiple disks attached to a processor (or SMP)

- **Disk blocks striped across disks**
- **Redundant disks can be used for reliability**

Network Attached Disks



Access disk(s) directly through network

- All nodes can access disk (s)
- No intermediate host

I/O Nodes

Access files through file server(s)

- Server(s) attached to other nodes through network

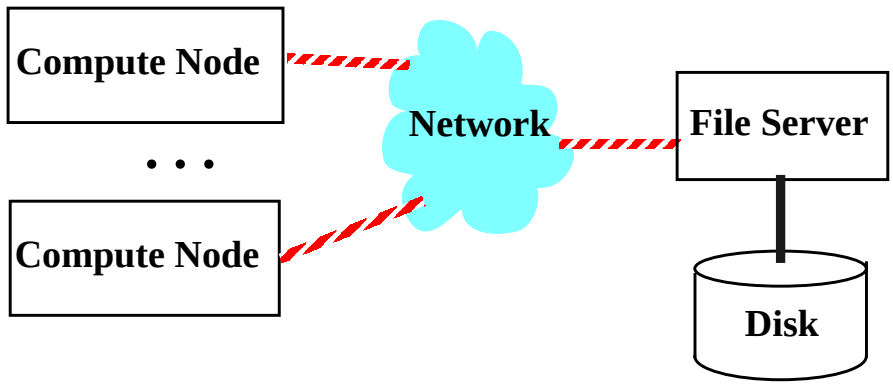
Workstation model

- Single file server (per filesystem)
- Distributed filesystem protocol, e.g., NFS, DFS.

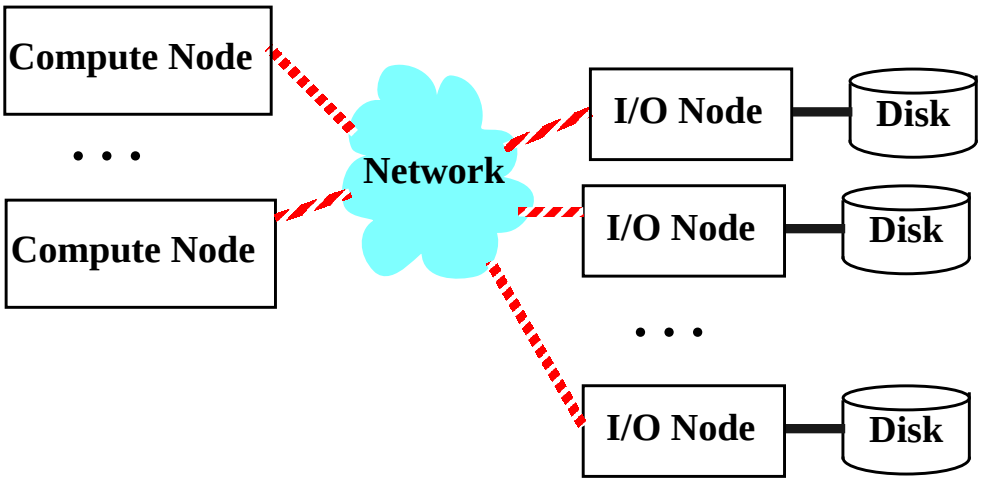
Parallel filesystem model

- Multiple file servers (per filesystem)
- Parallel filesystem protocol, e.g., PFS, PIOFS.

I/O Nodes - Workstation Model



I/O Nodes - Parallel Filesystem Model



Parallel I/O Hardware Summary

Disks

- **Disks can be single disks or hardware arrays**
- **Disks can be accessed directly or across a network**
- **Network accessed disks can be directly attached to the network or may require an I/O node**

Differences are largely logical

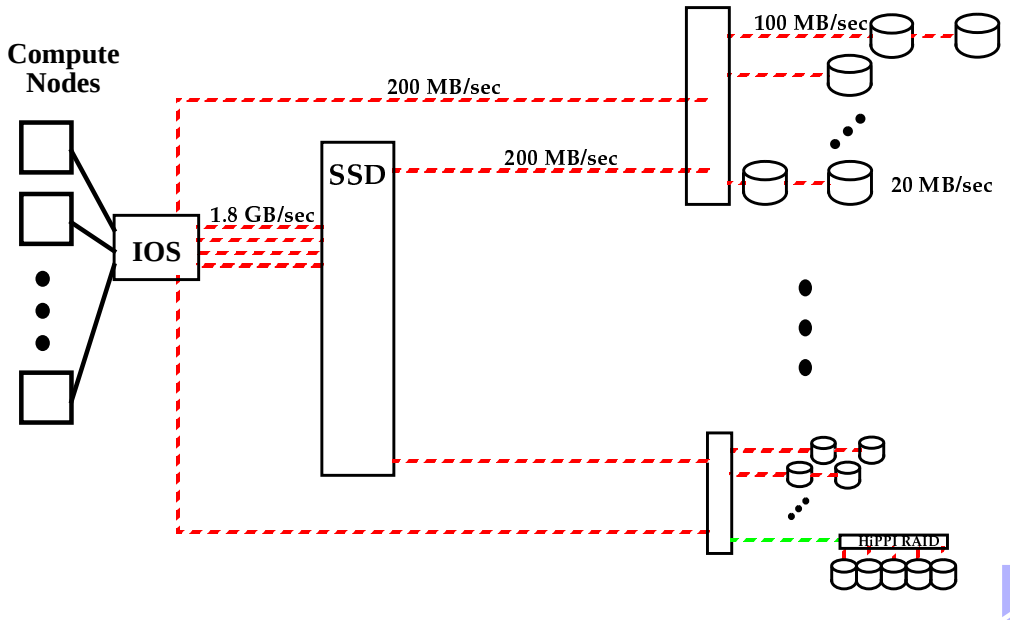
- **Network attached disks can be thought of as special I/O nodes**
- **SCSI channels can be thought of as networks**

Parallel I/O Hardware Summary

The real difference is the filesystem

- **Are files striped?**
  - **Internal to a hardware RAID**
  - **Across directly attached disks**
  - **Across I/O nodes**
- **Is a filesystem shared with other nodes?**

Cray Research C90 I/O Architecture



Cray Research C90

Shared memory compute nodes

- **Custom processors**
- **SMP OS, UNICOS**

Only a single system accesses each disk

- **Disks directly attached or network attached**
- **No I/O nodes or file servers**
- **Same as SMP workstation model**

Cray Research C90

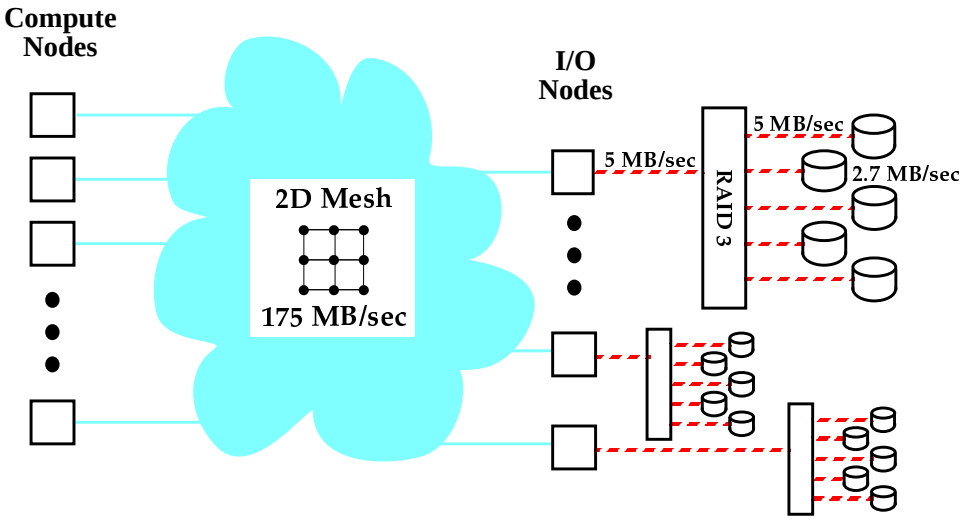
Can stripe across multiple disks

- software array (AED)

Can buffer through memory or solid state disk

I/O scales well with $$$

Intel Paragon I/O Architecture



Intel Paragon

Compute Nodes

- i860XP based
- Attached inside mesh (175 MB/sec)
- Distributed OS, OSF/1 AD

Disks attached only to I/O nodes

- Full compute nodes — Intel i860XP w/16-32 MB
- Attached inside mesh (175 MB/sec)
- 5 SCSI disks, Hardware RAID-3 (4+1)

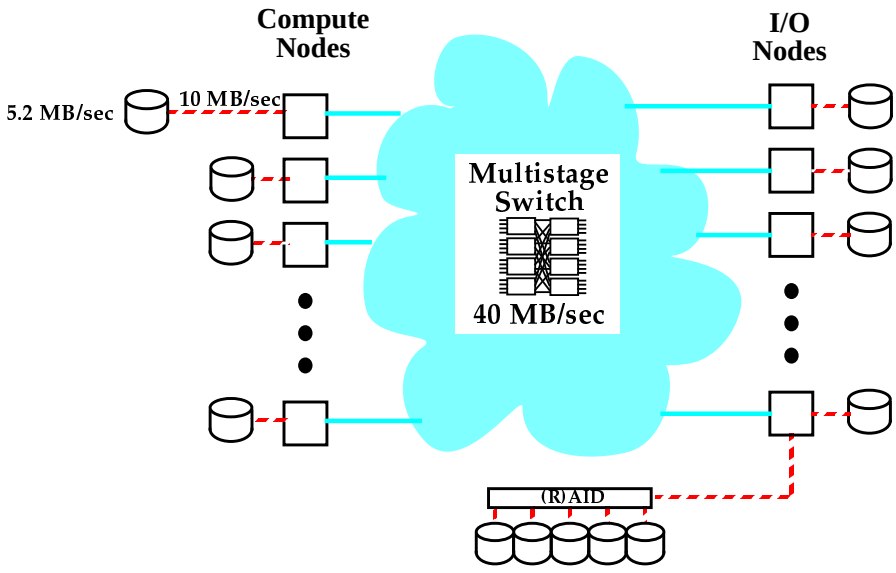
Intel Paragon

Parallel File System: *PFS*

- UNIX file interface, built on top of UFS
- File blocks striped across UFS filesystems

Performance scales at about 2.5MB/sec per I/O node

IBM SP2 I/O Architecture



IBM SP2

Compute Nodes

- Rack mounted RS6000 workstations
- Runs full AIX on every node

Disks directly attached to each node

- SCSI, SCSI-2, FCS, or any micro-channel adapter
- (R)AID or normal SCSI disks
- Nodes communicate through 40MB/s multistage network

IBM SP2

No clear differentiation between node types

- Every node has at least one disk, some have more
- I/O nodes defined by software
- I/O nodes can also be compute nodes

File systems

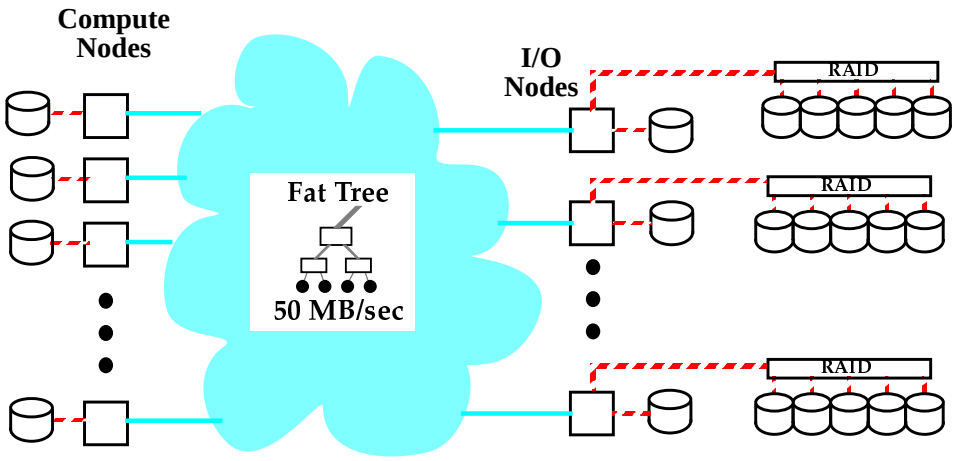
- Normal workstation distributed file systems (NFS, DFS, AFS, etc.)
- Parallel file system (PIOFS)

IBM SP2

Parallel file system: *PIOFS*

- UNIX interface, on top of IBM file system
- Files striped across I/O nodes' disk(s)
- Performance scales with both number of I/O nodes and disk speed to speed of internal network (10-15MB/sec for system messages)

Meiko CS-2 I/O Architecture



Meiko CS-2

Compute nodes

- SPARC based
- Full Solaris on every node

Disks directly attached to each node

- RAID or normal SCSI disks
- Every node has at least one disk
- Nodes communicate through 50MB/sec fat tree network

Meiko CS-2

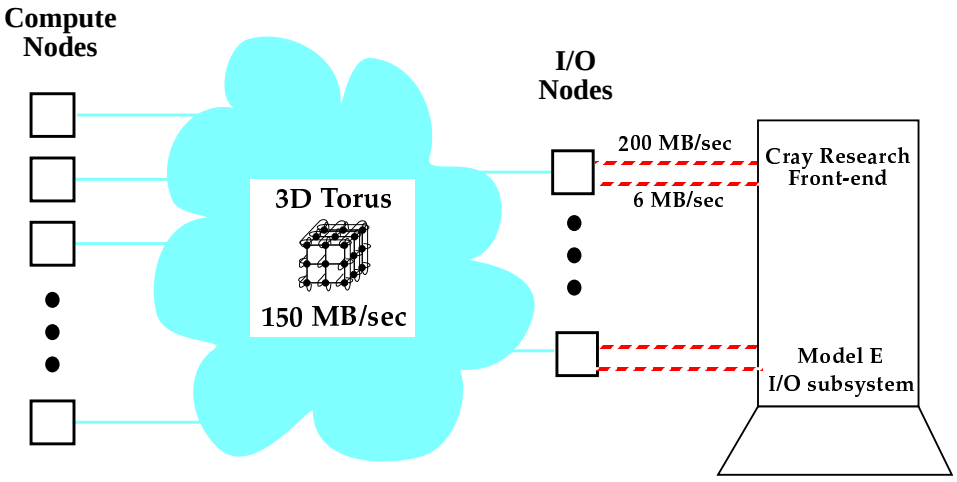
I/O nodes

- Same as compute nodes, but have a RAID
- Defined by software, can compute as well

Parallel file system

- Files striped across I/O node RAIDs
- Achieves about 2MB/s per I/O node

Cray T3D I/O Architecture



Cray T3D

Compute nodes

- DEC Alpha based
- Mach based micro-kernel on each node
- Front end services node requests

Disks attached to single file server: *Cray host*

- Centralized, UNIX file system on Cray host
- Disks directly attached to Y-MP (see C90 slide)
- Cray host SSD can be used as file cache

Cray T3D

I/O Nodes

- Modified compute nodes — DEC Alpha based
- Attached inside torus (X and Z only) (300 MB/sec)
- Attached to Y-MP through I/O Gateway

  1 HISP (200 MB/sec), 1 LOSP (6MB/sec)

I/O Hardware Example Summary

System	Compute Nodes	I/O Nodes	Disk Types	Disk Attachment	Striping File System	Performance
Cray C90	C90	n/a	Various	Cray IOS	Software AED	limited by $$
Intel Paragon	i860 Based	i860 Based	RAID-3	SCSI	PFS	~2.5MB/s per I/O node
IBM SP2	RS6000	RS6000	Various	SCSI, FCS, ...	PIOFS	up to 10-15 MB/s per I/O node
Meiko CS-2	SPARC	SPARC	SCSI Disks, RAID	SCSI	PFS?	~2MB/s per I/O node
Cray T3D	DEC Alpha	DEC Alpha	Various	I/O Gateway	Internal to Y-MP	limited by Y-MP